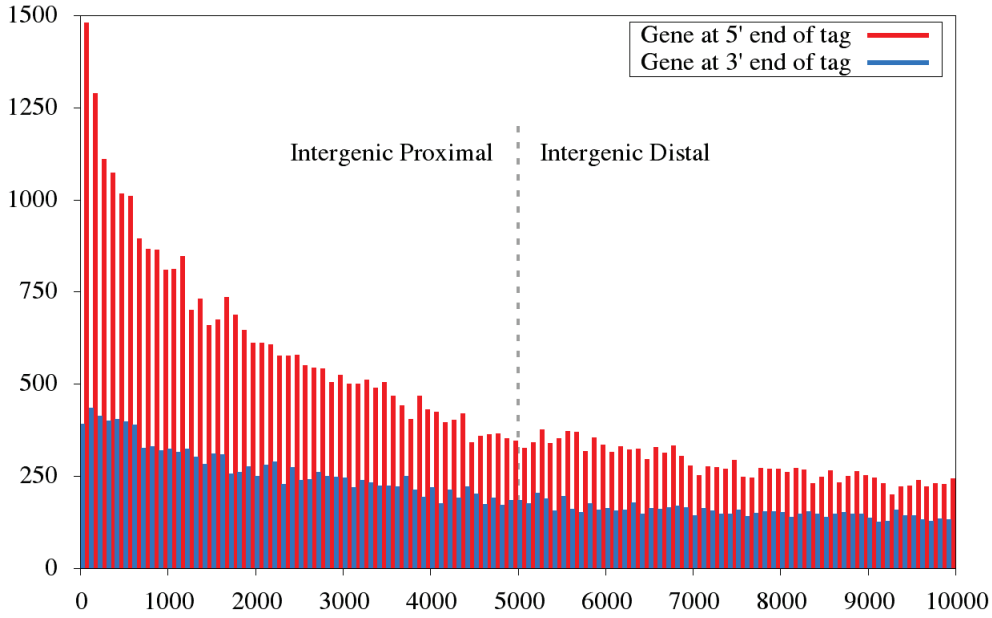
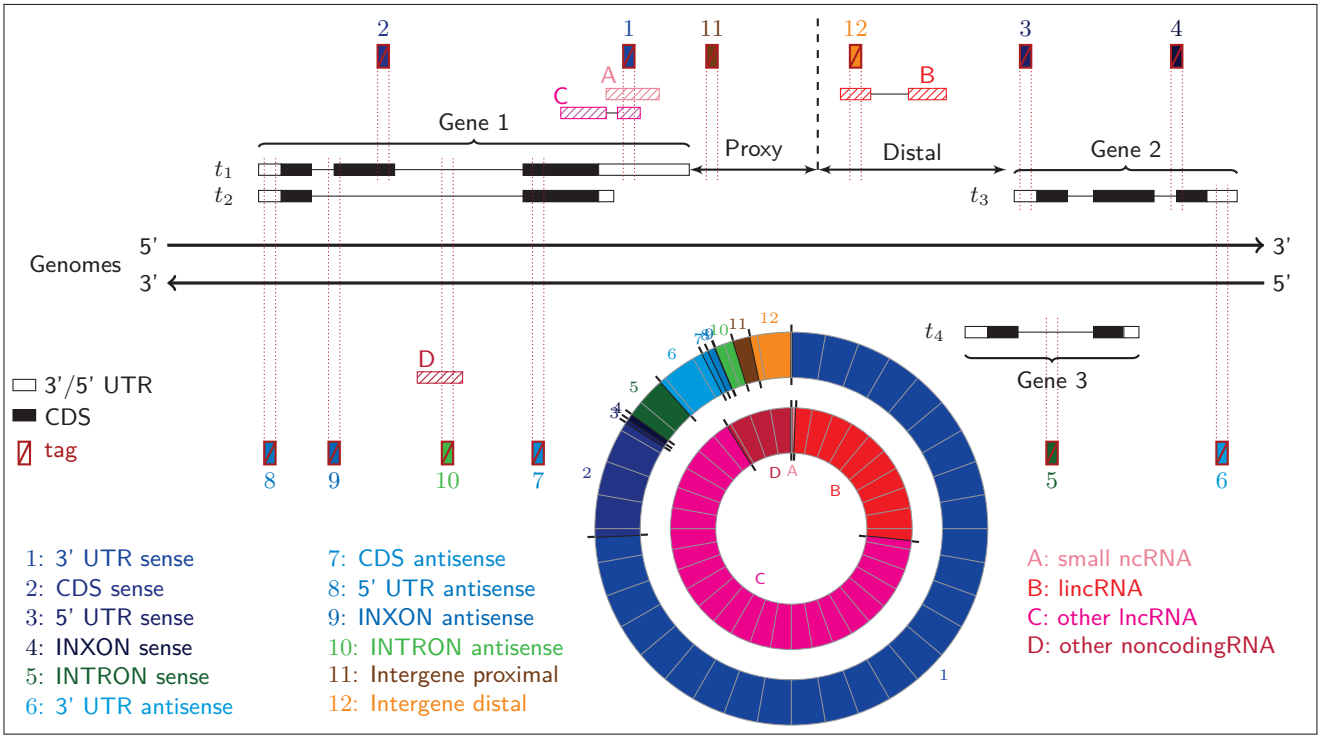
Supplemental Figure 1: Threshold between the intergenic distal and proximal regions



Distribution of tags relative to their distance from the 3' and 5' of the nearest gene. Each bar represents the number of tags according to their distance from the nearest gene (100bp intervals are shown). The number of tags neighbouring the 3' end of genes (blue bars) is stable at all distances analysed. Conversely, the number of tags neighbouring the 5' end of genes (red bars) is very high for tags in close proximity to the next gene and then it progressively decreases with the increase of the distance from the gene up to 5kbp where it becomes stable: this represents the threshold between the "Intergenic Proximal" and "Intergenic Distal" zones. The "Intergenic Proximal" is an ambiguous zone where new potential candidates and variants are indistinguishable, while "Intergenic Distal" represents a non-ambiguous zone of new candidate non-coding transcripts.

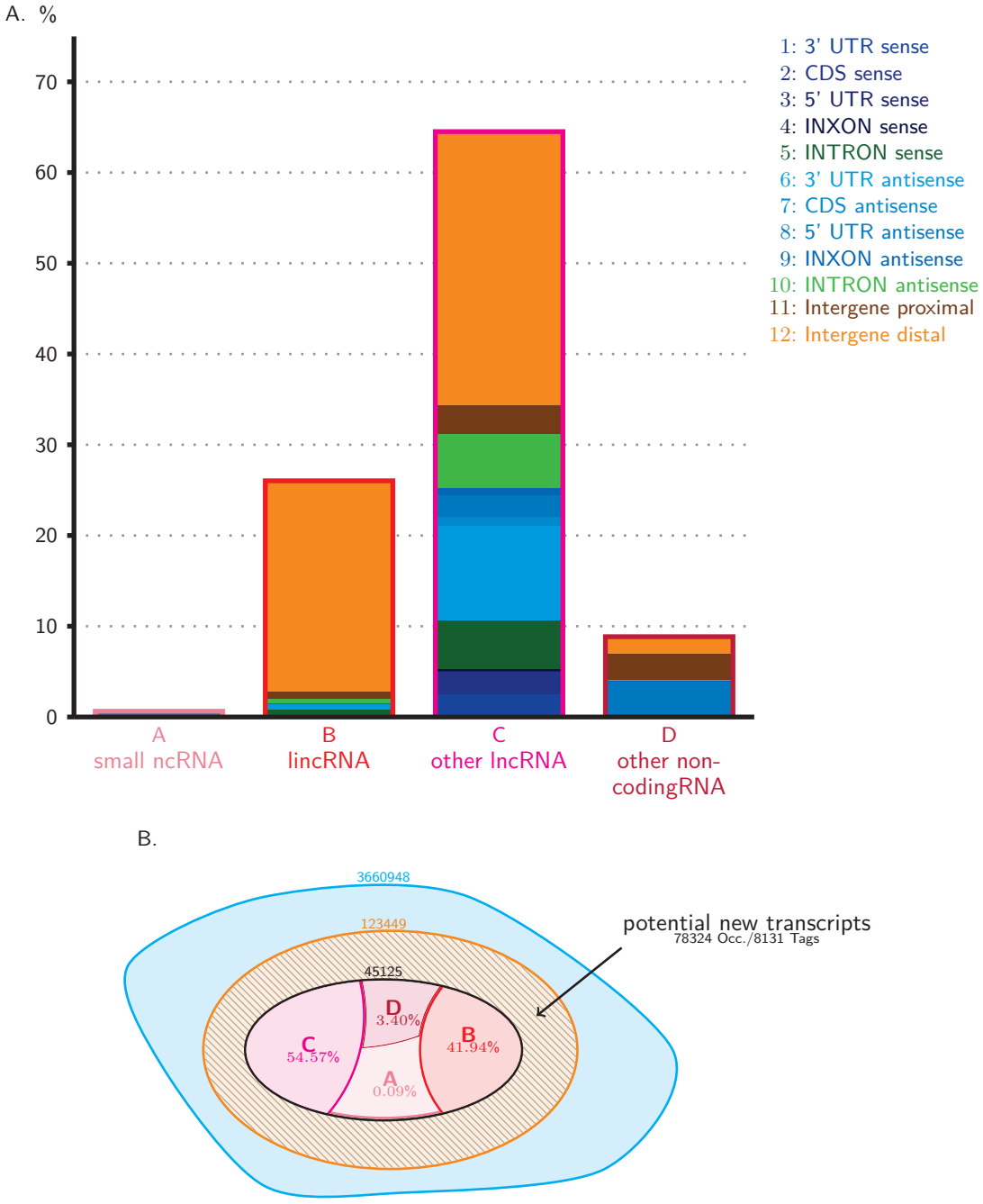
Supplemental Figure 2: DGE tag distribution in the AML-hs0430 library (tumour cells) according to Process A and Process B



We adopted a two-step strategy to annotate the subset of "LSAGE AML tumoral hs0430" tags that were uniquely mapped to the human genome using Ensembl: a general annotation process that considers protein-coding genes and pseudogenes (Process A) and a non-coding annotation process that considers only non-coding genes (Process B) (see Figure 1).

(Process A) Tags were classified with higher priority given to gene v/s intergenic annotations and strand information, as described in Material and Methods. A tag located in a gene (sense orientation) could be exonic (tag1, tag2, tag3) inxonic (tag4) or intronic (tag5). A tag located in a gene (but in the opposite strand) could be exonic (tag6, tag7, tag8), inxonic (tag9) or intronic (tag10). A tag outside a gene (intergenic localization) could be classified as proximal (tag11) or distal (tag12) to a 3' gene. The external pie chart indicates the genomic distribution of DGE sequences assigned to coding-genes based on the tag classification.

(Process B) Tags are classified according to their overlap with sequences of non-coding genes in: (A) small ncRNAs, (B) lincRNAs, (C) other lncRNAs and (D) other ncRNAs. A non-coding and a protein-coding gene could be identified by the same tag (e.g., tag 1 corresponds to the 3'UTR region of a coding transcript and to a non-coding transcript at the same time). In this case, we consider that the ncRNA transcript overlaps with a protein-coding gene. The internal pie chart shows the global genomic distribution of DGE sequences assigned to non-coding transcripts.



A. Bar chart describing the proportion of DGE sequences assigned to the different categories of non-coding transcripts in each genomic region. LincRNA and other lncRNA sequences are more abundant in the intergenic distal regions.
B. Pie chart showing the distribution of all "LSAGE AML tumoral hs0430" sequences in the human genome. The first inset pie chart represents all intergenic "LSAGE AML tumoral hs0430" sequences. The second inset pie chart corresponds to the four categories of non-coding transcripts in intergenic regions (small ncRNAs, lincRNAs, other lncRNAs and ncRNAs) with their relative proportions (%). The difference between the two pie charts represents orphan sequences without annotation, which could correspond to potential new transcripts or methodological artefacts. New non-annotated sequences are more abundant than non-coding sequences.

A. (Process A)

Type	Nb Occ.	(in %)	Nb Tags
1: 3' UTR sense	2721677	74.34	22150
2: CDS sense	356566	9.74	5584
3: 5' UTR sense	12276	0.34	368
4: INXON sense	17137	0.47	479
5: INTRON sense	129504	3.54	18170
6: 3' UTR antisense	149304	4.08	9798
7: CDS antisense	18931	0.52	2206
8: 5' UTR antisense	22218	0.61	1504
9: INXON antisense	1530	0.04	211
10: INTRON antisense	56397	1.54	7891
11: Intergene proximal	51959	1.42	3256
12: Intergene distal	123449	3.37	10071
Total	3660948	100.00	81688

B. (Process B)

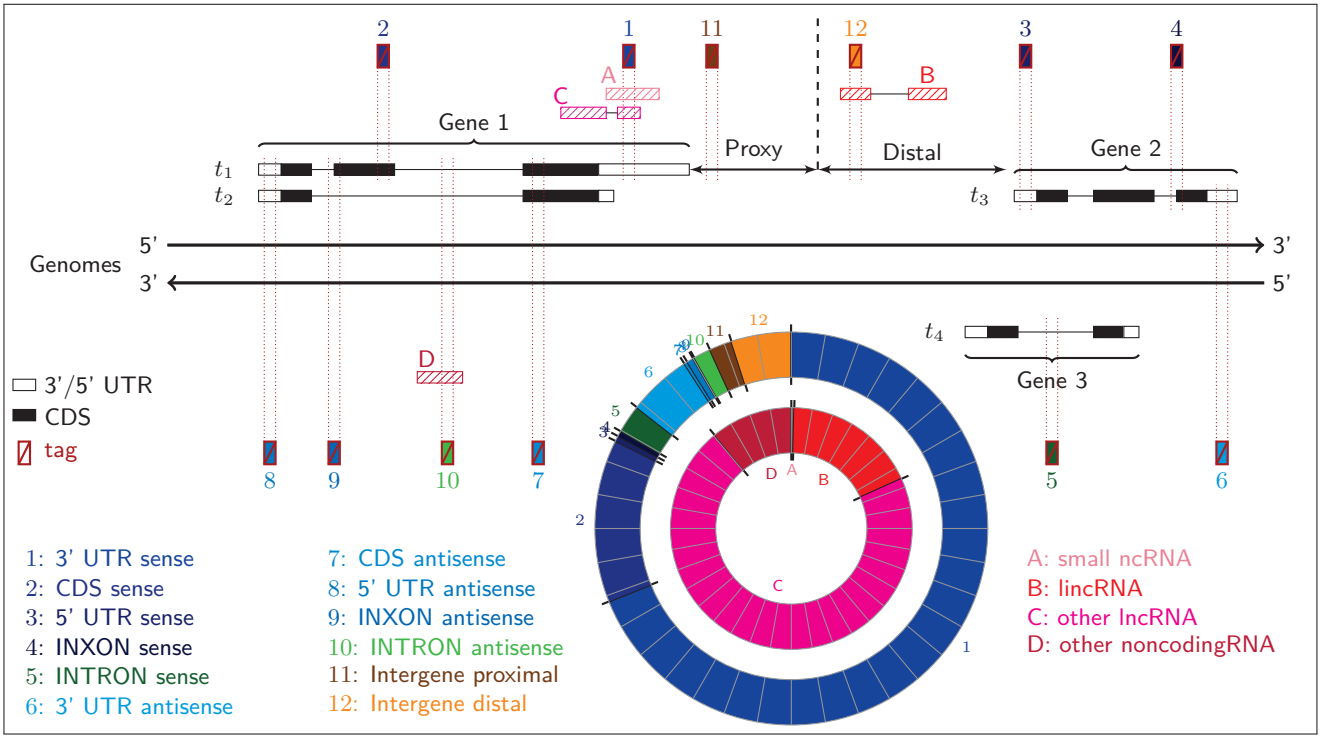
Type	Nb Occ.	(in %)	Nb Tags
A: small ncRNA	507	0.62	63
B: lincRNA	21259	26.04	690
C: other lncRNA	52668	64.51	3287
D: other noncodingRNA	7213	8.83	125
Total	81647	100.00	4165

C. (Process AxB)

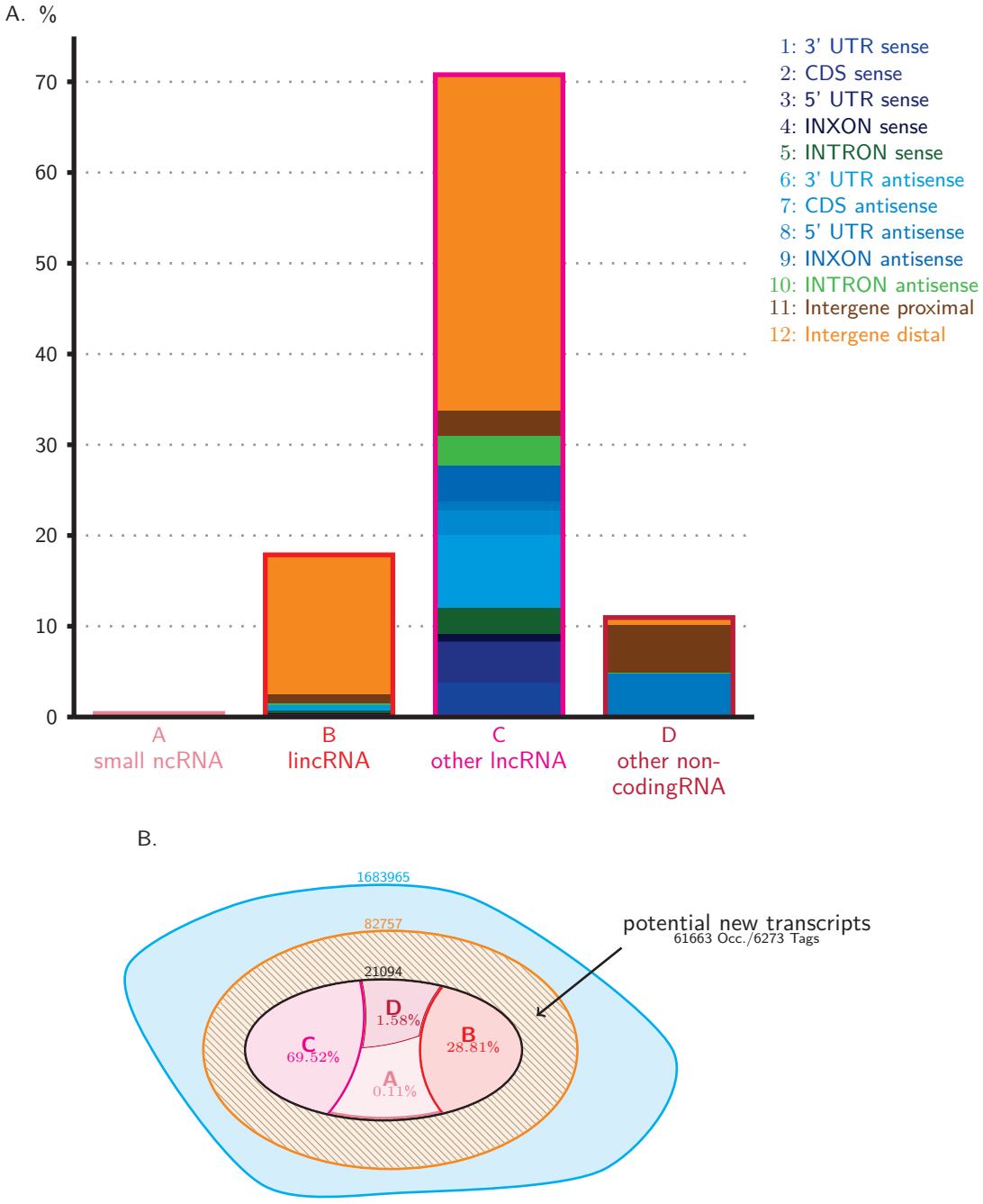
Class	A: small ncRNA	B: lincRNA	C: other lncRNA	D: other noncodingRNA	Total
1: 3' UTR sense	151	42	2076	23	2292
2: CDS sense	0	0	2011	2	2013
3: 5' UTR sense	6	0	54	0	60
4: INXON sense	0	48	148	0	196
5: INTRON sense	147	562	4426	14	5149
6: 3' UTR antisense	6	541	8464	58	9069
7: CDS antisense	5	16	865	0	886
8: 5' UTR antisense	5	40	1887	3143	5075
9: INXON antisense	0	2	695	5	702
10: INTRON antisense	29	363	4849	143	5384
11: Intergene proximal	116	718	2570	2292	5696
12: Intergene distal	42	18927	24623	1533	45125
Total	507	21259	52668	7213	81647

Detailed distribution, percentage and occurrences of "LSAGE AML hs0430" DGE tags with a unique match on the human genome. A. Genomic distribution and occurrences of DGE tags assigned to coding transcripts (Process A). B. Genomic distribution and occurrences of DGE tags assigned to non-coding transcripts (Process B). C. Global distribution and occurrences of DGE tags assigned to non-coding transcripts (Process AxB).

Supplemental Figure 3: DGE tag distribution in the hESC-hs0238 library according to Process A and Process B



We adopted a two-step strategy to annotate the subset of "LSAGE Embryonic stem cell hs0238" tags that were uniquely mapped to the human genome using Ensembl: a general annotation process that considers protein-coding genes and pseudogenes (Process A) and a non-coding annotation process that considers only non-coding genes (Process B) (see Figure 1).
(Process A) Tags are classified with higher priority given to gene v/s intergenic annotations and strand information, as described in Material and Methods. A tag located in a gene (sense orientation) could be exonic (tag1, tag2, tag3) inxonic (tag4) or intronic (tag5). A tag located in a gene (but on the opposite strand) could be exonic (tag6, tag7, tag8), inxonic (tag9) or intronic (tag10). A tag outside a gene (intergenic localization) could be classified as proximal (tag11) or distal (tag12) to a 3' gene. The external pie chart indicates the genomic distribution of DGE sequences assigned to coding-genes based on the tag classification.
(Process B) Tags were classified according to their overlap with sequences of non-coding genes in: (A) small ncRNAs, (B) lincRNAs, (C) other lncRNAs and (D) other ncRNAs. A non-coding and a protein-coding gene could be identified by the same tag (e.g., tag 1 corresponds to the 3'UTR region of a coding transcript and to a non-coding transcript at the same time). In this case, we consider that the ncRNA transcript overlaps with a protein-coding gene. The internal pie chart shows the global genomic distribution of DGE sequences assigned to non-coding transcripts.



A. Bar chart describing the proportion of DGE sequences assigned to the different categories of non-coding transcripts in each genomic region. LincRNA and other lncRNA sequences are more abundant in the intergenic distal regions.
B. Pie chart showing the distribution of all "LSAGE Embryonic stem cell hs0238" sequences in the human genome. The first inset pie chart represents all intergenic "LSAGE Embryonic stem cell hs0238" sequences. The second inset pie chart shows the four categories of non-coding transcripts in intergenic regions (small ncRNAs, lincRNAs, other lncRNAs and ncRNAs) with their relative proportions (%). The difference between the two pie charts represents orphan sequences without annotation, which could correspond to potential new transcripts or methodological artefacts. New non-annotated sequences are more abundant than non-coding sequences.

A. (Process A)

Type	Nb Occ.	(in %)	Nb Tags
1: 3' UTR sense	1160940	68.94	17357
2: CDS sense	225741	13.41	3628
3: 5' UTR sense	7457	0.44	140
4: INXON sense	7599	0.45	209
5: INTRON sense	39638	2.35	5881
6: 3' UTR antisense	88369	5.25	5453
7: CDS antisense	4614	0.27	551
8: 5' UTR antisense	9338	0.55	551
9: INXON antisense	2021	0.12	75
10: INTRON antisense	23953	1.42	2881
11: Intergene proximal	31538	1.87	2005
12: Intergene distal	82757	4.91	7594
Total	1683965	100.00	46325

B. (Process B)

Type	Nb Occ.	(in %)	Nb Tags
A: small ncRNA	156	0.39	25
B: lincRNA	7083	17.87	610
C: other lncRNA	28068	70.80	1743
D: other noncodingRNA	4337	10.94	73
Total	39644	100.00	2451

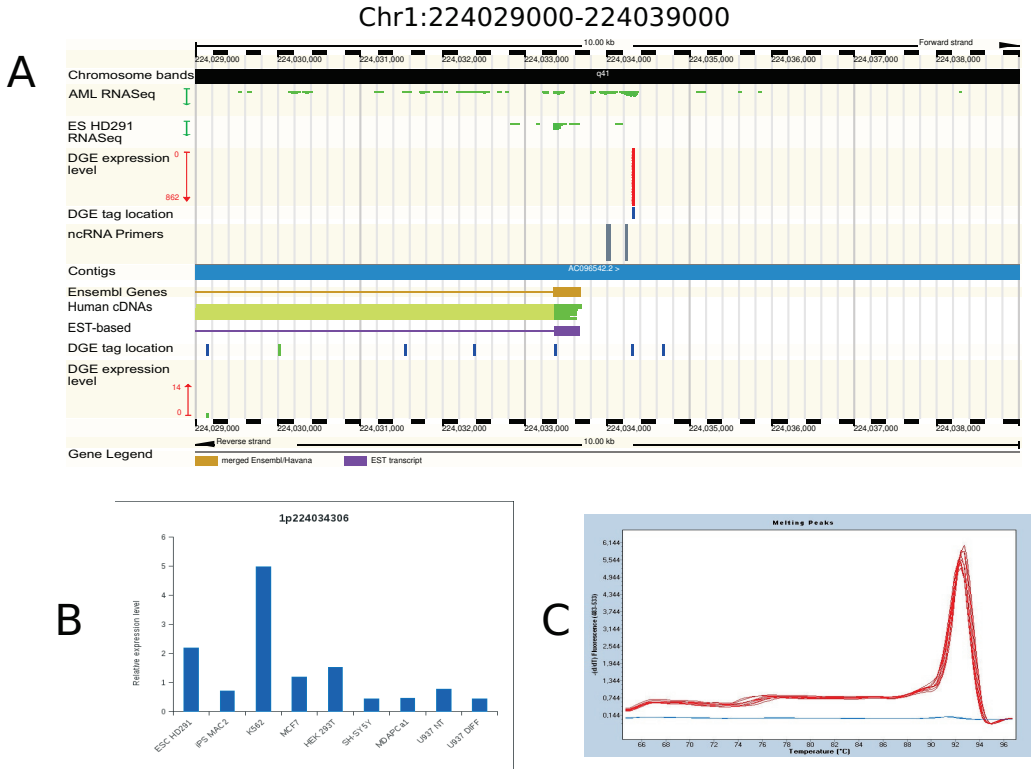
C. (Process AxB)

Class	A: small ncRNA	B: lincRNA	C: other lncRNA	D: other noncodingRNA	Total
1: 3' UTR sense	73	29	1524	82	1708
2: CDS sense	0	0	1753	2	1755
3: 5' UTR sense	0	0	56	0	56
4: INXON sense	0	161	294	0	455
5: INTRON sense	26	97	1138	5	1266
6: 3' UTR antisense	13	234	3211	28	3486
7: CDS antisense	0	22	1076	0	1098
8: 5' UTR antisense	0	2	361	1786	2149
9: INXON antisense	0	0	1603	3	1606
10: INTRON antisense	5	49	1266	35	1355
11: Intergene proximal	17	412	1123	2064	3616
12: Intergene distal	22	6077	14663	332	21094
Total	156	7083	28068	4337	39644

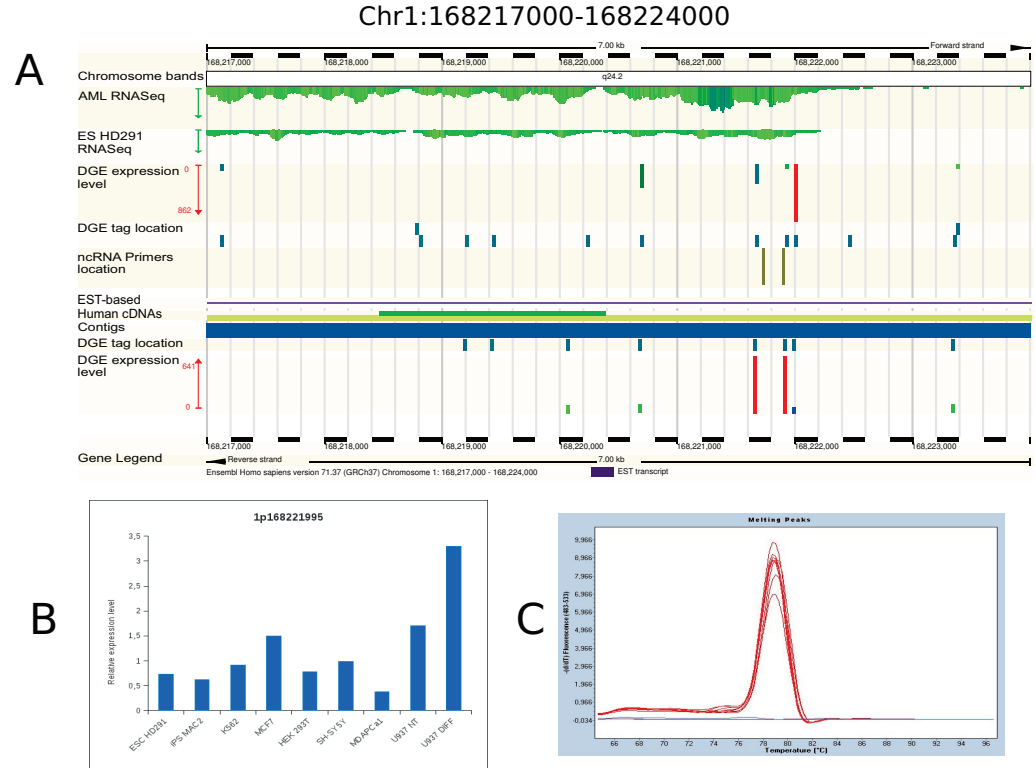
Detailed distribution, percentage and occurrences of "hESC-hs0238" DGE tags with a unique match on the human genome. A. Genomic distribution and occurrences of DGE tags assigned to coding transcripts (Process A). B. Genomic distribution and occurrences of DGE tags assigned to non-coding transcripts (Process B). C. Global distribution and occurrences of DGE tags assigned to non-coding transcripts (Process AxB).

Supplemental Figure 4: Other examples of new non-annotated transcripts

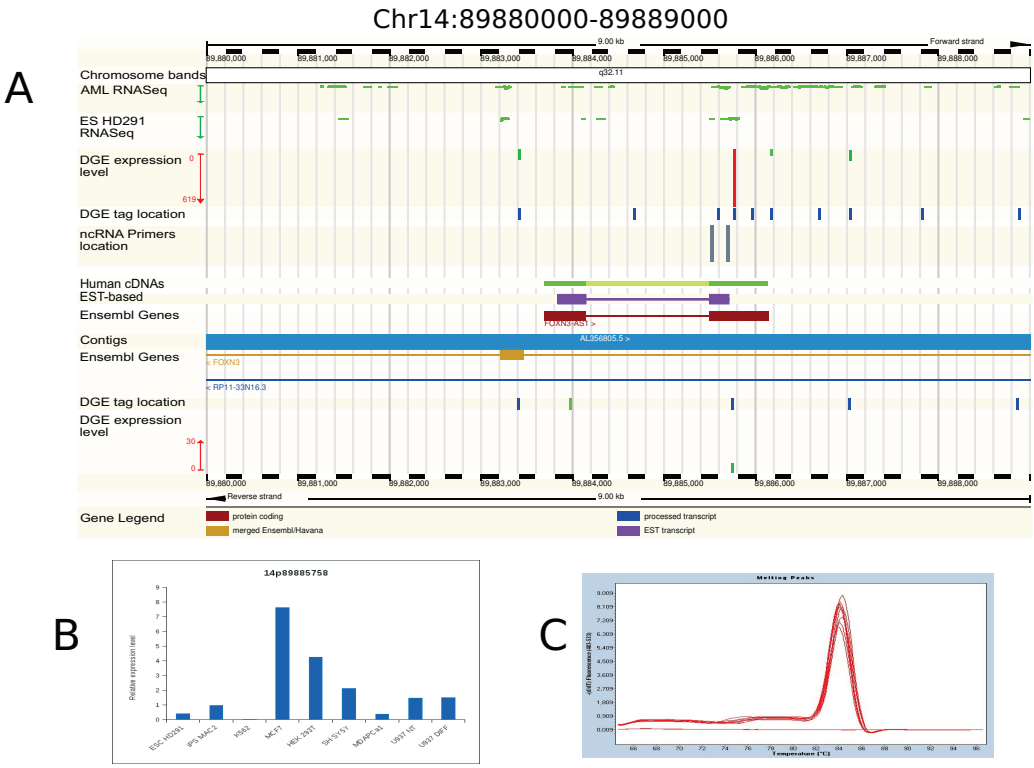
First example



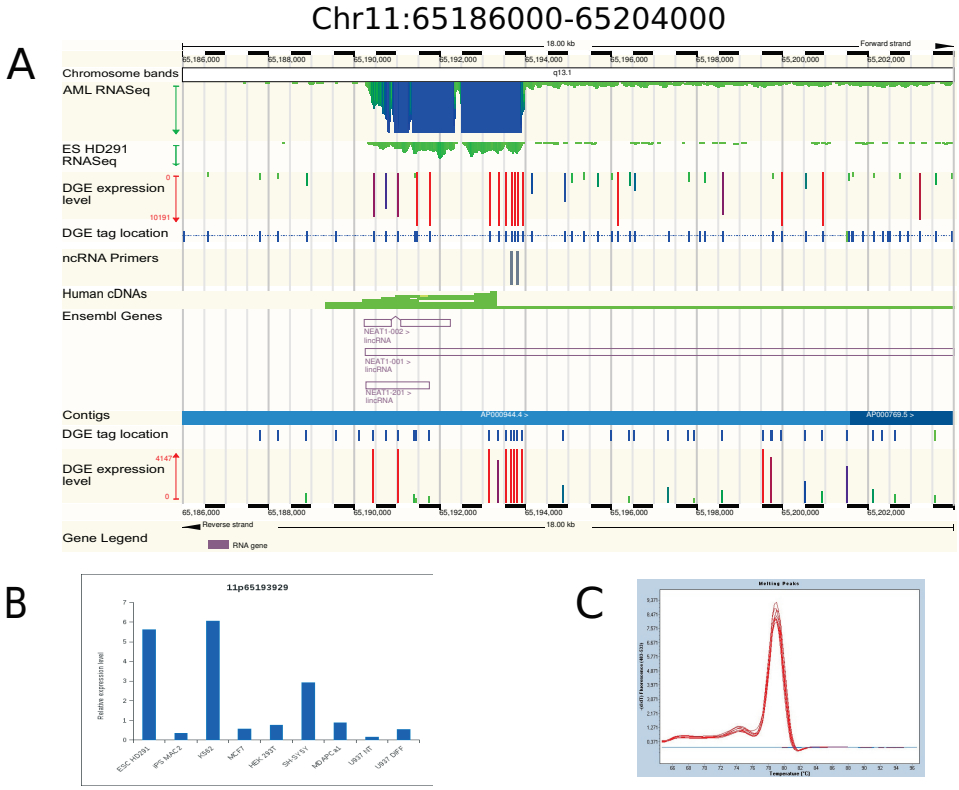
Second example



Third example

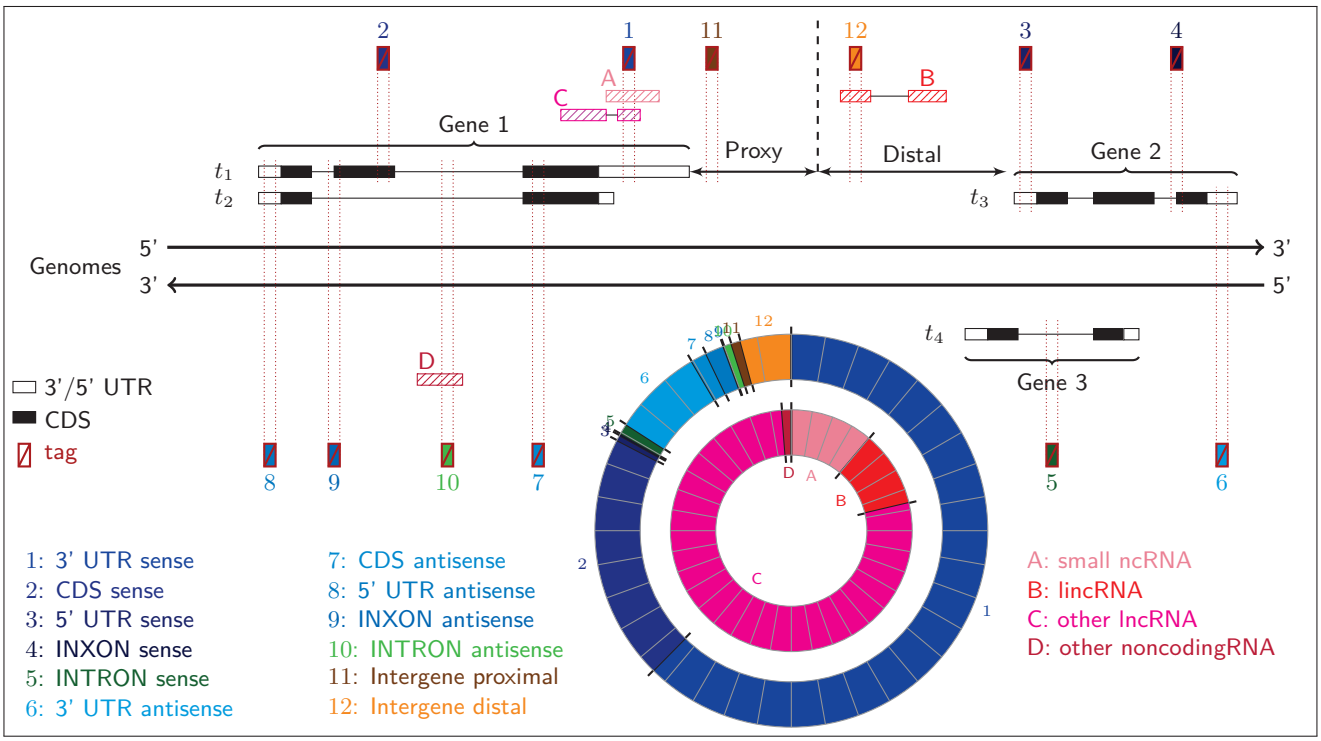


Fourth example



(A) Display of the Ensembl Genome Browser web page for new non-annotated transcripts. The blue horizontal bars represent chromosomes. Gene structures ("Ensembl Genes", "Human cDNAs", "EST-based" tracks) are annotated by Ensembl. Public and private DGE data ("DGE tag location" track: blue rectangle for occurrence ≥ 1, green for occurrence $= 1$) are displayed on both strands of the chromosome with their relative occurrences (histogram of the "DGE expression level" track) using a private DAS server. The histogram of RNA-Seq coverage (private data: RNA-Seq for hpSC and AML) in the chromosomal region is displayed on the top ("ES HD291 RNASeq " and "AML RNASeq" tracks). (B) Relative expression of new transcripts in different cell lines validated by quantitative Real Time PCR. (C) The corresponding melting curve analysis.

Supplemental Figure 5: Distribution of common human and mouse sequences



We adopted a two-step strategy to annotate the subset of "common human and mouse sequences" tags that were uniquely mapped on the human genome using Ensembl: a general annotation process that considers protein-coding genes and pseudogenes (Process A) and a non-coding annotation process that considers only non-coding genes (Process B) (see Figure 1).
(Process A) Tags were classified with higher priority given to gene v/s intergenic annotations and strand information, as described in Material and Methods. A tag located in a gene (sense orientation) could be exonic (tag1, tag2, tag3) inxonic (tag4) or intronic (tag5). A tag located in a gene (but on the opposite strand) could be exonic (tag6, tag7, tag8), inxonic (tag9) or intronic (tag10). A tag outside a gene (intergenic localization) could be classified as proximal (tag11) or distal (tag12) to a 3' gene. The external pie chart indicates the genomic distribution of DGE sequences assigned to coding-genes transcripts based on the tag classification.
(Process B) Tags were classified according to their overlap with sequences of non-coding genes in: (A) small ncRNAs, (B) lincRNAs, (C) other lncRNAs and (D) other ncRNAs. A non-coding and a protein-coding gene could be identified by the same tag (e.g., tag 1 corresponds to the 3'UTR region of a coding transcript and to a non-coding transcript at the same time). In this case, we consider that the ncRNA transcript overlaps with a protein-coding gene. The internal pie chart shows the global genomic distribution of DGE sequences assigned to non-coding transcripts.

A. (Process A)

Type	Nb Occ.	(in %)	Nb Tags
1: 3' UTR sense	3971026	62.40	1429
2: CDS sense	1286639	20.22	2762
3: 5' UTR sense	28621	0.45	335
4: INXON sense	9814	0.15	63
5: INTRON sense	56009	0.88	454
6: 3' UTR antisense	473688	7.44	985
7: CDS antisense	83718	1.32	1430
8: 5' UTR antisense	96959	1.52	253
9: INXON antisense	4573	0.07	53
10: INTRON antisense	39040	0.61	306
11: Intergene proximal	49077	0.77	82
12: Intergene distal	264432	4.16	553
Total	6363596	100.00	8705

B. (Process B)

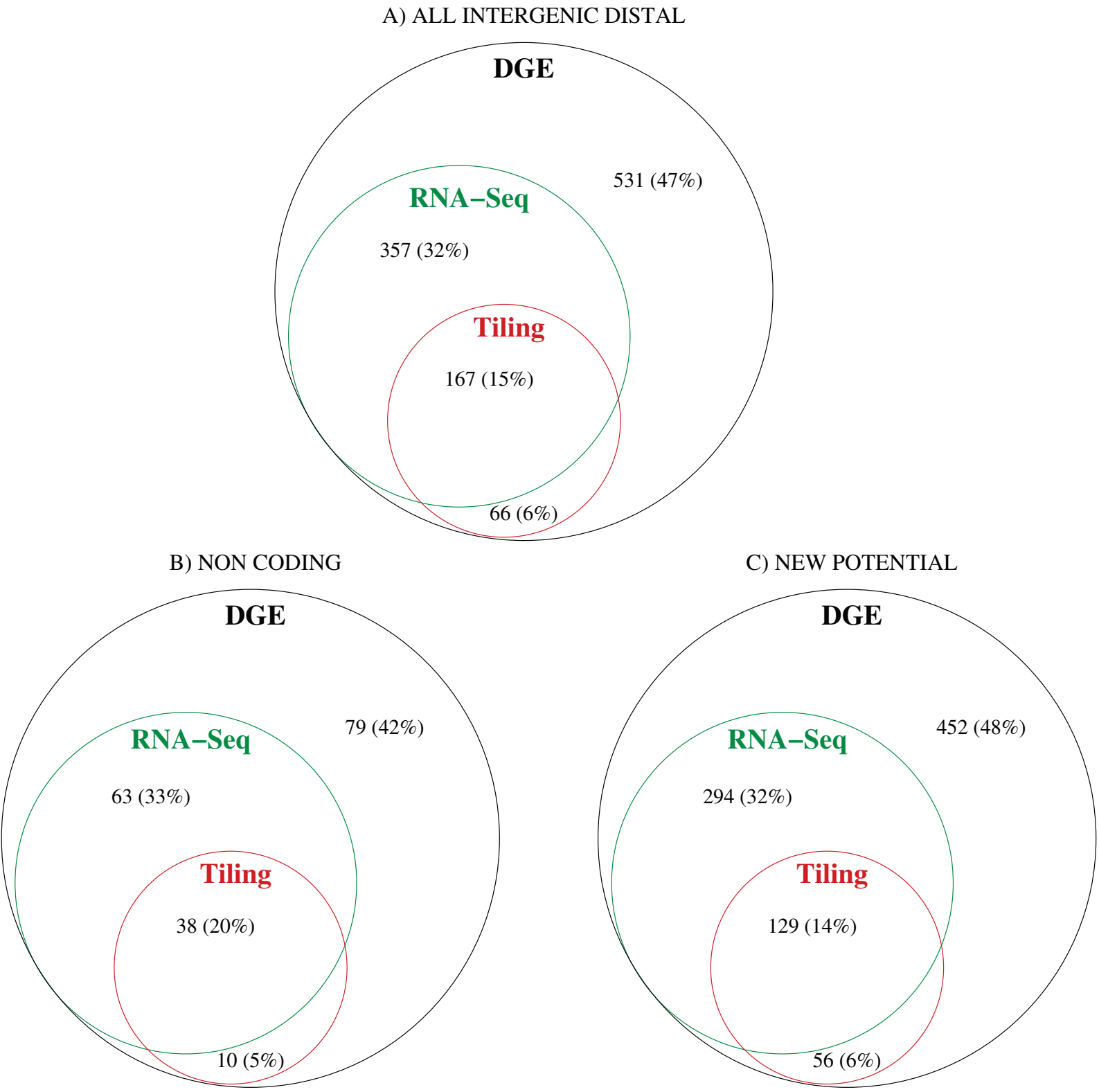
Type	Nb Occ.	(in %)	Nb Tags
A: small ncRNA	14889	11.16	17
B: lincRNA	13562	10.16	40
C: other lncRNA	103370	77.45	222
D: other noncodingRNA	1644	1.23	14
Total	133465	100.00	293

C. (Process AxB)

Type	Nb Occ.	(in %)	Nb Tags
1: 3' UTR sense	1404	1.05	13
2: CDS sense	44122	33.06	14
3: 5' UTR sense	9	0.01	2
4: INXON sense	28	0.02	2
5: INTRON sense	2156	1.62	19
6: 3' UTR antisense	25596	19.18	65
7: CDS antisense	3455	2.59	65
8: 5' UTR antisense	4202	3.15	26
9: INXON antisense	0	0	0
10: INTRON antisense	547	0.41	12
11: Intergene proximal	2156	1.62	8
12: Intergene distal	49790	37.31	67
Total	133465	100	293

Detailed distribution, percentages and occurrences of common "TranscriRef" DGE tags with a unique match on the human and mouse genomes. A. Genomic distribution and occurrences of DGE tags assigned to coding transcripts (Process A). B. Genomic distribution and occurrences of DGE tags assigned to non-coding transcripts (Process B). C. Global distribution and occurrences of DGE tags assigned to non-coding transcripts (Process AxB).

Supplemental Figure 6: Venn diagram of the 1,121 intergenic tags that are highly expressed in hESC



Venn diagrams showing the distribution of the 1,121 "TranscriRef" tags that are highly expressed in hESCs as indicated by the clustering analysis (Figure 5). The coverage of each tag by the DGE, RNA-Seq and tiling arrays methods is given and then a value is computed when applying the filtering process of the digitagCT pipeline to qualify a tag as specifically expressed or not. To select specific candidate transcripts for a tissue, a "TranscriRef" tag was retained only if i/ the DGE tag had occnb >= 2, ii/ was covered by three RNASeq reads overlapping the 5' of the tag region and iii/ it hybridized in Tiling arrays. A) Distribution of the 1,121 intergenic distal tags: 524 tags (47%) were covered by both DGE and RNA-Seq; 167 of these common tags (15%) were also covered by tiling arrays. B) Distribution of the 190 intergenic tags already annotated as non-coding transcripts by the Ensembl Genome Browser. C) Distribution of tags that could correspond to new potential transcribed regions.

Supplemental Table 1: Description of all librairies used in “TranscriRef”

see Table-S1-List-DGE-Libraries.xls

Supplemental Table 2: Gene Onthology analysis of conserved human and mouse tags

see Table S2-GO-chart-DAVID-DB-EXON-INTRON-ANTISENSE.xls

Supplemental Table 3: Biological validation of 36 new potential transcripts

see Table-S3-List-36-candidates.xls